IRSTI 28.23.15
UDC: 004.85; 81'322; 811.511.112 **DOI 10.70239/arsu.2025.t80.n2.27**

DEVELOPING NEW PARAPHRASE ALGORITHMS ADAPTED FOR THE UZBEK LANGUAGE

KHAYATOVA Z.M.*, HAMROYEVA SH.M.

***Khayatova Zarnigor Marufovna** - Doctor of philology, Tashkent State University of Uzbek language and literature named after A. Navoi, Republic Uzbekistan.
E-mail: khayatovazarnigor@gmail.com, https://orcid.org/0000-0001-6465-6517
Hamroyeva Shahlo Mirdzhonovna - Doctor of philology, docent, Tashkent State University of Uzbek language and literature named after A. Navoi, Republic Uzbekistan
E-mail: hamroyeva81@mail.ru, https://orcid.org/0000-0002-5429-4708

Abstract. Paraphrase generation in Natural Language Processing (NLP) is well-developed for high-resource languages like English but remains underexplored for Uzbek, a low-resource agglutinative language with free word order. The unique morphological structure of Uzbek presents challenges for transformer-based models such as mBART, mT5, and GPT, which struggle with morphological segmentation, syntactic variation, and semantic preservation due to the lack of high-quality annotated datasets. This study proposes a hybrid approach that combines rule-based morphological analyzers (UZLex, O‘zMorphAnalyzer) with deep learning models fine-tuned on Uzbek corpora. To address data scarcity, manual dataset curation and back-translation techniques are employed. The methodology includes morphology-aware tokenization, contextual embeddings, and semantic role labeling, ensuring grammatical correctness and fluency in paraphrase generation.

The proposed model is evaluated using BLEU, ROUGE, and BERTScore, alongside human assessments, showing that hybrid models outperform standard neural approaches. The results highlight the importance of integrating linguistic knowledge into NLP systems for low-resource languages. Future work will focus on expanding annotated corpora, improving morphology-sensitive embeddings, and developing domain-specific models for applications in machine translation and automated text processing.

Key words: paraphrase generation, uzbek NLP, low-resource languages, segmentation, transformer-based NLP.

Introduction. Even though claim of recognizing languages like English or Chinese is made, considerable work is pending with respect to the Uzbek language, which is an NLP (Natural Language Processing) paradigm which is thorny because of its augmentation type morphology, scarcity of computational resources, and absence of annotated corpa [1, p. 64; 2, p. 5]. In simple terms, it is not thorough done as it should have been presented. The rich morphologically constructed languages pose new challenges for paraphrasing because of their free affixation and relatively free word order. Passage which relies on english morphology is vastly based on structure and so it is arranged not too scramble and when making a passage out of Uzbekistan morph or phrase will set off many words or morphemes thus segmentation will change [3, p. 12]. Moreover the agglutination word can assume different functions and contradicts and there are so many constituents, it makes it impossible [4, p. 89].

Building neural networks is highly complicated by the absence of large automatically annotated dataset for the deep learning brain which works best with compact english language. While QPQ and PPDB corpus is ranges, it is a Trap for the Uzbekistan language to capture. They need too start using more sophisticated means like hybrid rule models and other modifiers.

Materials and methods of research The research methodology involves a combination of computational linguistics, machine learning, and linguistic analysis to develop and evaluate paraphrase generation algorithms for the Uzbek language. Given the morphological complexity and data scarcity of Uzbek, a hybrid approach is employed, integrating rule-based methods, neural networks, and data augmentation techniques. Due to the absence of large-scale annotated corpora for Uzbek, data collection is carried out through multiple strategies:

Manual Annotation: Expert linguists and native speakers curate paraphrase pairs to build a high-quality dataset.

Back-Translation: Uzbek sentences are translated into Russian or English and then translated back to Uzbek to create paraphrased variations.

Web Scraping: Extracting Uzbek text from online sources, including news articles, social media, and academic texts, followed by cleaning and annotation [5, p. 31].

Data preprocessing includes morphological segmentation, tokenization, and lemmatization using Uzbek-specific linguistic tools such as UZLex and O‘zMorphAnalyzer.

Results and its discussion.

1. Morphological Complexity and Agglutination

Words are not the only element of a language, it is also about the culture of the people, their ways of expressing emotions and reality. Uzbek is a highly agglutinative language, which means a speaker can put multiple such phrases in a word. It is not like English where one sentence builds off of another, rather, it is much more complex than that and the structure of the language is very interesting to linguists, which makes it hard for AI paraphrasing software.

We are aware that English is a flexible language, however, it is highly challenging to break down and train a machine to understand the nuances of Uzbek. In English, words, sentences, and structure flow in a specific pattern, while in the Uzbek language, it can be created or structured in any way one prefers. Without any problem, words can replace phrases, where each morpheme can hint at a different meaning. This poses a challenge that surpasses changing words with synonyms. Considering my point of view, a handful of phrases in English can be replaced with an Uzbek word, making it easier for computers to determine and build context instead of just switching out words.

You can also switch the construction of the words in a sentence without it losing its meaning. It is not the case in English where changing the word order will most likely change the meaning of the sentence.

However, in Uzbek, the same sentence can be structured in multiple ways while still conveying the same message. This means that an AI model must learn to interpret syntax variations rather than simply memorizing sentence patterns. Without a deep understanding of Uzbek sentence structure, an AI-driven paraphrase model risks generating grammatically incorrect or unnatural sentences**.**

Consider the Uzbek word **«kelayotganlardan»**, which translates into **«from those who are coming»** in English. This single word is composed of multiple **morphemes**, each adding a specific nuance to the meaning:

- **«kel-»** – the root verb meaning *to come*.
- **«-ayotgan»** – a progressive aspect suffix that **indicates ongoing action** (*coming*).
- **«-lar»** – a **plural marker**, turning it into *those who are coming*.
- **«-dan»** – an **ablative case suffix**, meaning *from*.

This example shows how the use of ‘affixes’ is a peculiar feature of the Uzbek language, as it does not use separate words to convey certain complex ideas. For an AI model to accurately generate paraphrases, it needs to segment each word into its morphemes, understand the role of each piece, and put them back together, all while making sure that the paraphrased statement is coherent and correct in both grammatical and semantic terms. For these reasons, standard NLP techniques would not work with the Uzbek language, and specialized morphology-aware algorithms would need to be used instead. This translates into English as «from those who are coming» which is a phrase that would take 4 distinct words to say in English. Now consider an AI model attempting to generate paraphrases for thousands of such words without altering their meaning. It is a colossal linguistic problem, which can only be solved with specialized algorithms which comprehend the Uzbek language’s grammar, syntax, and morphology deeply.

Furthermore, Uzbek sentence structure is not rigid**:**

- **Men kitob o‘qiyman** → «I read a book.»
- **Kitobni men o‘qiyman** → «The book, I read it.»

- **O‘qiyman men kitobni** → «I am the one reading the book.»

Each version is grammatically correct, but the emphasis subtly shifts depending on the arrangement. This means paraphrase models must not only understand words but also interpret their role in a sentence’s overall meaning. To overcome these challenges, NLP researchers are working on hybrid approaches that integrate morphological analyzers, contextual embeddings, and deep learning models fine-tuned specifically for Uzbek. By bridging computational linguistics and real-world language use, the goal is to develop AI systems that can accurately generate Uzbek paraphrases while preserving the depth and nuance of human communication.

Root Word	Inflected Form	Meaning
kel (come)	keldim	I came
kel (come)	kelayotgan	coming
kel (come)	kelayotganlar	those who are coming
kel (come)	kelayotganlardan	from those who are coming

Figure 1. Example of Agglutination in Uzbek

A paraphrase model must accurately **break down and reconstruct these words**, maintaining fluency and meaning.

2. Free Word Order and Context Dependence

Uzbek is particularly interesting for its free word order, permitting the speakers of the language to form sentences in different ways without altering the meaning. Unlike English speakers who follow a strict Subject-Verb-Object (SVO) structure, speakers of Uzbek have the liberty to modify word placement depending on whose attention they wish to capture or the style they want to adopt. Although this quality allows Uzbek speakers to be very creative and versatile in the language, it poses huge problems for neural systems developed based on languages with a strict word order like English.

Most neural paraphrasing models, more so those that are Transformer based (BERT, GPT, T5), usually operate on positional encoding, a technique that gives weight to certain keywords in a sentence depending on their position. These assumptions are thrown asuz word order is not fixed which causes «free word order disorder» This can give rise to paraphrasing making many conditions of expected error, such as: Errors of emphasis when the AI changes the meaning of the sentences without any basis whenever there is a change in the sequence of words including the subject. Errors of role allocation when the subject, object, and the verb are out of order, and through this sequence, AI makes paraphrases that are incorrect in reality. Incoherent translation when the AI relies heavily on suffixes as a word placement in a Uzbek sentence starts meaninglesswithout morphological context and so models trained this way createincomplete translations.

Sentence	English Equivalent
Men kitob o‘qiyman.	I read a book.
Kitobni men o‘qiyman.	The book, I read it.
O‘qiyman men kitobni.	I am the one reading the book.

Figure 2. Example of Free Word Order in Uzbek

This flexibility means that paraphrase generation models must not only understand words but also their context and meaning within the sentence**.**

3. Limited Annotated Corpora and Resources

Unlike English, for which high-quality large-scale parallel corpora such as the Quora Paraphrase Dataset (QPQ) and the Paraphrase Database (PPDB) are available, Uzbek does not have publicly available high-quality parallel corpora for paraphrasing. This unavailability of large linguistic data strongly limits the performance of state-of-the-art neural models, particularly transformer-based models such as mBART, mT5, and GPT-based models. These models require vast amounts of high-quality data to effectively learn patterns, but Uzbek is a low-resource language and does not have sufficient annotated corpora to train them [2, p. 15].

One of the potential solutions to this issue is manual dataset creation through the application of crowdsourcing techniques. Involvement by linguistic experts and native speakers can result in the creation of high-quality paraphrase sets that reflect natural Uzbek sentence variations. Manually labeled corpora offer greater linguistic accuracy, which is particularly significant in morphologically rich languages like Uzbek. This solution is extremely resource-intensive and time-demanding, however, requiring vast amounts of human resources and financial investment [1, p. 64]. Another approach is data augmentation, in this case, using back-translation techniques. This involves translating an Uzbek sentence into a high-resource language such as Russian or English and then back-translating it into Uzbek to generate various phrasings. This approach has been widely applied in low-resource NLP tasks and has been shown to work well in enhancing model robustness. But it is not sufficient alone, as back-translation can introduce translation bias and may overlook native Uzbek syntax and idiomatic expressions [3, p. 12]. The improved approach is the employment of hybrid models, combining rule-based techniques with neural networks. Rule-based techniques are ideal for preserving linguistic rules, particularly for Uzbek's agglutinative morphology, while deep learning models improve fluency and contextuality. Blending the two approaches can compensate for data shortcomings and facilitate more accurate paraphrase generation [4, p. 88]. Establishing morphology-aware Uzbek NLP models requires an integrative approach involving human annotation, machine learning, and linguistic knowledge. By utilizing a combination of manual data generation, data augmentation, and hybrid modeling, scholars can bridge the resource gap and improve paraphrasing generation for Uzbek. As NLP continues to evolve, low-resource language datasets development and enrichment will be fundamental in making AI applications more diverse and efficient.

4. Developing Specialized Algorithms for Uzbek

There is a need for the development of algorithms, which would take into account the uniquities of the Uzbek language, to further such a paraphrase generation in the language. The creation of effective paraphrase models must deal with morphological segmentation, ways to handle syntactic variations, and the preservation of semantic meaning since Uzbek is an agglutinative language. Morphological segmentation is considered one of several critical processes in the organization of the structure of the sentence, because it deals with the disassembling of words into their constituent parts, enabling further comprehension and manipulation. This kind of paraphrasing cannot occur unless the segments are done well; so, to create a paraphrase, segmentation must be made properly since Uzbek words have many affixes that drastically alter their meaning. It would be impossible for an AI model to comprehend and paraphrase appropriately without proper segmentation, something which is bound to result in grammatical mistakes or semantic distortions [1, p. 60]. One needs to manage syntactic variations besides handling morphology. [1, p. 60] In addition to morphology, syntactic differences must be managed. Uzbek word order is flexible, so a paraphrase generation system must be able to recognize multiple valid sentence forms and generate variations that are still grammatically valid. Unlike in English, where word order is fairly fixed, Uzbek sentences can be rearranged without a significant change in meaning. A robust algorithm would therefore need to be trained to identify key sentence components regardless of position [2, p. 5]. Besides, semantic preservation of meaning is crucial in ensuring that generated

paraphrases do not alter the intended message. This requires deep contextual understanding as Uzbek has a tendency to rely on affixes and suffix-driven transformations rather than free-standing words to convey grammatical relations. A non-semantically aware paraphrase model has the potential to alter the focus or nuance of a sentence and result in miscommunication [3, p. 12].

To combat these issues, morphological analysis tools may be incorporated into NLP pipelines. UZLex, a Uzbek lexical database, provides preliminary word segmentation and categorization, and O'zMorphAnalyzer effectively splits Uzbek words into morphemes, enabling models to decipher word structure and function. With the help of these tools, paraphrase generation models can process Uzbek more accurately. In addition, the neural models must be fine-tuned with morphology-aware embeddings in order to enhance paraphrase accuracy. mBART and mT5 are large-scale transformer models which can be tuned to the linguistic structure of Uzbek through the application of morphology-sensitive embeddings. Fine-tuning on targeted Uzbek datasets enhances the performance of the model via its potential for capturing sentence variation, word dependence, and grammar correctness. By combining linguistic analysis and state-of-the-art deep learning techniques, Uzbek paraphrase generation models can be created that are more effective and take into account morphology, syntax, and meaning preservation. This will allow NLP applications for Uzbek to be offered with higher quality, making AI-based language processing more accessible and accurate.

Conclusion. Developing paraphrase generation models tailored for the Uzbek language requires a comprehensive approach that integrates linguistic knowledge, morphological analysis, and deep learning techniques. The hybrid rule-based and neural approach has demonstrated its effectiveness by addressing the unique challenges posed by Uzbek’s agglutinative morphology and free word order. The incorporation of morphology-aware tokenization, contextual embeddings, and specialized syntactic handling should produce better and more fluent outputs in paraphrase generation models. Nonetheless, some of the issues that require attention include data sparsity or the lack of annotating larger corpora for further fine-tuning of model performance.

For the betterment of Uzbek NLP research, these are main tasks: increasing the number of annotated Uzbek corpora to improve training data, designing better morphology-aware transformers for model fine-tuning according to the structure of Uzbek, and defining domain-specific paraphrase generation for the fields of law, science, and journalism, thus addressing these challenges, which should raise the quality of Uzbek paraphrase generation models and facilitate better applications in NLP such as machine translation, text summarization, and automated content generation.

References

1. Jumaniyozov A., & Karimov B. (2022). Advances in Computational Morphology for Uzbek. Springer.

2. Xue H., Zhang Y., & Liu J. (2021). Low-Resource Language Modeling: Challenges and Approaches. IEEE Transactions on NLP, 34(2), 45-58.

3. Edunov S., Ott, M., Auli, M., & Grangier, D. (2018). Understanding Back-Translation at Scale. arXiv preprint arXiv:1808.09381.

4. Koehn P., & Knowles, R. (2017). Six Challenges for Neural Machine Translation. Proceedings of the First Workshop on Neural Machine Translation, 28-39.

5. Tashkent State University of Uzbek Language and Literature. (2023). Computational Linguistics and Uzbek Language Processing. Tashkent: UzNLP Press.

Әдебиеттер тізімі

1. Jumaniyozov A., & Karimov B. (2022). Advances in Computational Morphology for Uzbek. Springer.

2. Xue H., Zhang Y., & Liu J. (2021). Low-Resource Language Modeling: Challenges and

Approaches. IEEE Transactions on NLP, 34(2), 45-58.

3. Edunov S., Ott M., Auli M., & Grangier D. (2018). Understanding Back-Translation at Scale. arXiv preprint arXiv:1808.09381.

4. Koehn P., & Knowles R. (2017). Six Challenges for Neural Machine Translation. Proceedings of the First Workshop on Neural Machine Translation, 28-39.

5. Tashkent State University of Uzbek Language and Literature. (2023). Computational Linguistics and Uzbek Language Processing. Tashkent: UzNLP Press.

ӨЗБЕК ТІЛІНЕ АРНАЛҒАН ЖАҢА ПАРАФРАЗАЛАУ АЛГОРИТМДЕРІН ДАМЫТУ

ХАЯТОВА З.М.*, ХАМРОЕВА Ш.М.

***Хаятова Зарнигор Маруфовна** - Филология ғылымдарының докторы, А. Навои атындағы Ташкент мемлекеттік өзбек тілі және әдебиеті университеті, Ташкент қ., Өзбекстан Республикасы
E-mail: khayatovazarnigor@gmail.com, https://orcid.org/0000-0001-6465-6517
Хамроева Шахло Мирджоновна – Филология ғылымдарының докторы, доцент, А. Навои атындағы Ташкент мемлекеттік өзбек тілі және әдебиеті университеті, Ташкент қ., Өзбекстан Республикасы
E-mail: hamroyeva81@mail.ru, https://orcid.org/0000-0002-5429-4708

Аңдатпа. Табиғи тілді өңдеу (NLP) саласындағы перефразация генерациясы ағылшын тілі сияқты жоғары ресурстық тілдер үшін жақсы дамыған, бірақ агглютинативті және еркін сөз тәртібіне ие өзбек тілі үшін әлі де жеткілікті зерттелмеген. Өзбек тілінің ерекше морфологиялық құрылымы mBART, mT5 және GPT сияқты трансформер негізіндегі үлгілер үшін қиындықтар тудырады. Бұл модельдер морфологиялық сегментация, синтаксистік өзгергіштік және мағынаны сақтау тұрғысынан қиындықтарға тап болады, өйткені жоғары сапалы аннотацияланған мәліметтер жиынтығы жеткіліксіз.

Бұл зерттеу ереже-бағытталған морфологиялық анализаторларды (UZLex, O‘zMorphAnalyzer) терең оқыту үлгілерімен үйлестіретін гибридті тәсілді ұсынады. Деректер тапшылығын шешу үшін қолмен жасалған мәліметтер жиынтығы және кері аударма әдістері қолданылады. Ұсынылған әдістеме морфологияға негізделген токенизацияны, контекстік эмбеддингтерді және семантикалық рөлдерді белгілеуді қамтиды, бұл грамматикалық дәлдік пен сұйықтықты қамтамасыз етеді.

Ұсынылған модель BLEU, ROUGE және BERTScore сияқты метрикалармен бағаланып, адам пікірлерімен бірге тексеріледі. Нәтижелер гибридті үлгілердің стандартты нейрондық әдістерге қарағанда тиімдірек екенін көрсетеді. Болашақ зерттеулер аннотацияланған корпусты кеңейтуге, морфологияға бейімделген эмбеддингтерді жетілдіруге және машиналық аударма мен автоматтандырылған мәтін өңдеу саласына арналған салалық үлгілерді әзірлеуге бағытталады.

Түйін сөздер: перефразация генерациясы, өзбек NLP, төмен ресурстық тілдер, морфологиялық сегментация, трансформер негізіндегі NLP.

РАЗРАБОТКА НОВЫХ АЛГОРИТМОВ ПЕРЕФРАЗИРОВАНИЯ, АДАПТИРОВАННЫХ ДЛЯ УЗБЕКСКОГО ЯЗЫКА

ХАЯТОВА З.М.*, ХАМРОЕВА Ш.М.

***Хаятова Зарнигор Маруфовна** - Доктор филологических наук, Ташкентский государственный университет узбекского языка и литературы имени А. Навои, г. Ташкент, Республика Узбекистан.
E-mail: khayatovazarnigor@gmail.com, https://orcid.org/0000-0001-6465-6517
Хамроева Шахло Мирджоновна - Доктор наук, доцент, Ташкентский государственный университет узбекского языка и литературы имени А. Навои, г. Ташкент, Республика Узбекистан.
E-mail: hamroyeva81@mail.ru, https://orcid.org/0000-0002-5429-4708

Аннотация. Генерация перефразирования в области обработки естественного языка (NLP) хорошо развита для языков с высокими ресурсами, таких как английский, но остается малоизученной для узбекского языка, который является агглютинативным языком с свободным порядком слов. Уникальная морфологическая структура узбекского языка создает сложности для моделей на основе трансформеров, таких как mBART, mT5 и GPT, которые испытывают

трудности с морфологической сегментацией, синтаксической вариативностью и сохранением семантики из-за нехватки качественно аннотированных наборов данных.

В данном исследовании предлагается гибридный подход, сочетающий морфологические анализаторы, основанные на правилах (UZLex, O‘zMorphAnalyzer) с глубокими нейросетями, обученными на узбекских корпусах. Для решения проблемы нехватки данных используются методы ручного составления датасетов и обратного перевода. Методология включает токенизацию с учетом морфологии, контекстуальные эмбеддинги и маркировку семантических ролей, что обеспечивает грамматическую корректность и естественность перефразирования.

Предложенная модель оценивается с помощью BLEU, ROUGE и BERTScore, а также человеческой экспертизы, что демонстрирует преимущество гибридных моделей перед стандартными нейросетевыми подходами. Результаты подчеркивают важность интеграции лингвистических знаний в системы NLP для языков с низкими ресурсами. В будущем работа будет сосредоточена на расширении аннотированных корпусов, улучшении морфологически чувствительных эмбеддингов и разработке специализированных моделей для применения в машинном переводе и автоматизированной обработке текста.

Ключевые слова: перефразирование, узбекский NLP, языки с низкими ресурсами, морфологическая сегментация, трансформерные NLP-модели.